Efficient Visual Odometry and Mapping for Unmanned Aerial Vehicle Using ARM-based Stereo Vision Pre-Processing System

Changhong Fu , Adrian Carrio , Pascual Campoy

Fig. 1: Asctec Pelican quadrotor platform equipped with our newly designed light small-scale low-cost stereo vision pre-processing system.

I. INTRODUCTION

In the past decades, the visual odometry and mapping algorithms have been researched and developed fruitfully in the robot community, they can provide 6 Degrees-Of-Freedom (DOF) pose estimation and useful environment (obstacle) information for robots to autonomously navigate/explore in different types of GPS-denied cluttered environments.

In the literature, monocular and stereo cameras are widely applied as two main vision tools for visual odometry and mapping. For monocular camera, M. Pizzoli et al [1] et al recently presented a real-time probabilistic monocular pose estimation method for 3D dense reconstruction. And J. Engel et al [2] proposed a direct monocular Simultaneous Localization and Mapping (SLAM) algorithm for building consistent maps of the environments. These works have achieved promissing results, however, a monocular camera cannot sufficiently estimate the real absolute scale to the surrounding environments, which generates a lot of accumulated scale drifts for visual odmetry, especially in large-scale environments. Although many other works have alleviated this kind of problem by fusing other extra sensors, e.g. J. Zhang et al [3] adopted a 3D lidar device (i.e. a motor actuated rotated Hokuyo UTM-30LX) to enhance the visual odometry performance of a monocular camera. Our former work [4] fused an IMU sensor and a monocular camera onboard a UAV to successfully finish a see-and-avoid task during visual inspection, but the performances in all these works mainly lie on the measurement accuracy of those extra sensors, and the higher performance/quality of those extra sensors will result in more expensive systems. And besides, some of the extra sensors are still too heavy to be carried onboard a typical small-scale/multi-rotor UAV, as the Asctec Pelican Quadrotor[1] shown in Fig.1, and require more computational capability from the onboard computer of the UAV.

Stereo cameras can effectively estimate the disparity/depth information (i.e. scale) determined by the baseline between the left and the right camera, thereby improving the visual odometry and mapping performances. However, stereo cameras also have two bottlenecks: (I) when the distance between the robot and the target is much larger than the baseline, the depth estimation becomes inaccurate or simply invalid. (II) the features seen by only one side camera (e.g. occlusion) cannot be associated with the depth via real-time stereo matching, but these 2D features can provide useful information to strengthen the visual pose estimation. Most works in this field have configured a large baseline to solve the first limitation in large-scale environments, e.g. A. Geiger et al [5] set the baseline to more than a half meter for autonomy exploration, but this approach is not suitable for those typical UAVs.

Recently, different kinds of small-scale embedded system-based standalone stereo vision devices have been designed and applied for robots to process visual information, e.g. VisLab 3DV system [6], Skybotix VI sensor [7] and DLR's stereo device [8]. However, some of them are too heavy for UAVs, e.g. VisLab 3DV: 550 grams. And some of them have a high cost, e.g. Skybotix VI sensor: 3900 Euros. Furthermore, the embedded systems used in some of those devices have $dual$-core ARM[2] for parallelly processing onboard tasks, e.g. VisLab 3DV system and Skybotix VI sensor. Additionally, some of them are only demonstrated and utilized for $dense$ visual odometry tasks. All these limitations reduce the number of potential university/company end-users for a wide variety of UAV applications.

Nonetheless, the main contributions of this work are listed below:

- A new light small-scale low-cost ARM-based stereo vision pre-processing system for typical UAV has been designed, as the details introduced in Section II, which has advantages in terms of size, weight, cost and computational performance.
- A software plugin has been implemented on this newly designed system for robust and efficient stereo visual odometry application, which effectively takes advantage of 2D (e.g. far/occluded features without depth) and 3D (e.g. near features with depth) information, thereby solving the limitation of a small fixed baseline.
- Applied this system for real flights of a typical UAV, as discussed in Section V.

The structure of the paper is organized as follows: Section II introduces the details of our newly designed stereo vision pre-processing system. The software plugin used in this new system for visual odometry and 3D mapping are described in Section III. Section IV presents the performance of this specifically developed software plugin and one 3D mapping result. In Section V, one of results during real UAV flights is presented. Finally, conclusion and future work are proposed in Section VI.

[2]http://www.arm.com/

II. STEREO VISION SYSTEM

Nowadays, UAVs are being widely applied in civilian applications, e.g. disaster rescue, orchard monitoring, building fault inspection. For a typical small-scale/multi-rotor UAV, its size, payload, computation capability and expanded mounting space for other sensors are limited, e.g. LinkQuad[3], Asctec Pelican/Firefly, DJI F450/F550[4] (diagonal wheelbase: $\sim$50cm, without propellers, the maximum payload: 300-650 grams), therefore, light small-scale onboard pre-processing systems designed for UAV applications have became popular recently, allowing to save enough computing capability for the host computer onboard UAV to process other onboard tasks, e.g. controller, path planning and sensor fusion.

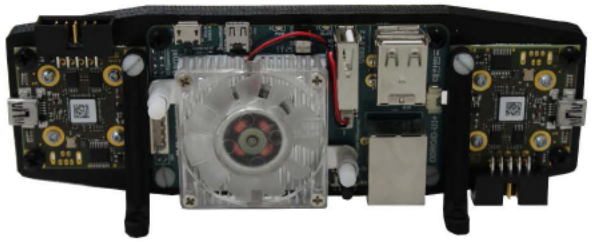
Considering size, weight, cost, computation performance, mounting flexibility and the capability to process information from complex surrounding environments, we have designed a new stereo vision pre-processing system based on an ARM processor, as shown in the Fig. 2, the details of our newly designed system are listed below:

- $computer$: it is the modification of hardkernel ODROID U3[5] ($69), which has one 1.7 GHz (manufactured clock frequency) $Quad$-Core processor (i.e. Samsung Exynos 4412 Prime Cortex-A9), 2 GByte RAM, 64 GByte eMMC-based storage space ($79), 10/100 Mbps Ethernet with RJ-45 LAN Jack, 3 High speed USB2.0 Host ports, 1 micro HDMI, 1 micro USB, and GPIO/UART/I2C ports. Its size is 83mm$\times$48mm, the weight is 48g (including heat sink). And power supply is 5V DC. In our current stereo vision system, the operating system is Ubuntu 13.04/13.10/14.04, it supports with the Hydro/Indigo version of Robot Operating System (ROS)[6], and OpenCV library[7] is also compatible. In addition, it also supports a wireless communication module.
- $cameras$: the system is equipped with two IDS uEye industry cameras[8] (type: UI-1221LE-C-HQ) based on CMOS type sensors (model: MT9V032C12STC) with USB 2.0. The camera supports High Dynamic Range (HDR) mode and global shutter. The frame rate reaches up to 87.2 FPS with freerun mode. In our stereo vision system, two uEye cameras are parallelly fixed on the two sides of a light multi-function mechanical part (which is also used to flexibly mount on the UAVs), and the stereo image pairs are synchronized with hardware trigger, their maximum image resolutions are 752$\times$480 pixels. The focal length of the lenses (i.e. Lensagon[9] BM2820) is 2.8mm, the horizontal and vertical fields of view are 98^o and 73^o, respectively. Each camera size is 36.0mm$\times$36.0mm$\times$25.2mm, and the weight is 20 grams.

[3]http://www.uastech.com/
[4]http://www.dji.com/
[5]http://www.hardkernel.com/
[6]http://www.ros.org
[7]http://www.opencv.org/
[8]http://en.ids-imaging.com/
[9]http://www.lensation.de/

The total weight of whole system is 100 grams, which is lighter than other frequently-used sensors, e.g. RGB-D sensor (Asus Xtion Pro Live): ∼200 grams, 2D Laser (Hokuyo UTM30-LX): ∼270 grams. And it also has less weight than other embedded system-based pre-processing stereo vision devices, as mentioned in the Section I. The dimension is 160mm×55mm×40mm, and its baseline is 12 centimeters. Additionally, the cost of our stereo camera system is only 800 Euros. To authors's best knowledge, this is the first work to present such a new light small-scale low-cost ARM-based stereo vision pre-processing system.

(a) Front view.



(b) Back view.

Fig. 2: Our newly designed light small-scale low-cost stereo vision pre-processing system.

III. VISUAL ODOMETRY AND MAPPING

This section describes the stereo visual odometry and mapping algorithm, the coordinate system and whole software architecture are shown in Fig. 4 below.

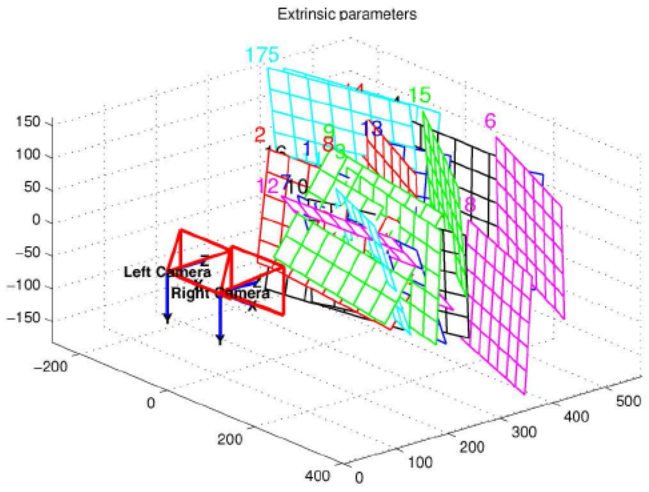


Fig. 3: One example of stereo calibration result (Unit: mm).

A. Disparity/Depth Map Estimation

The cameras in our system are calibrated using Camera Calibration Toolbox[10]. Fig. 3 shows one example of stereo calibration result.

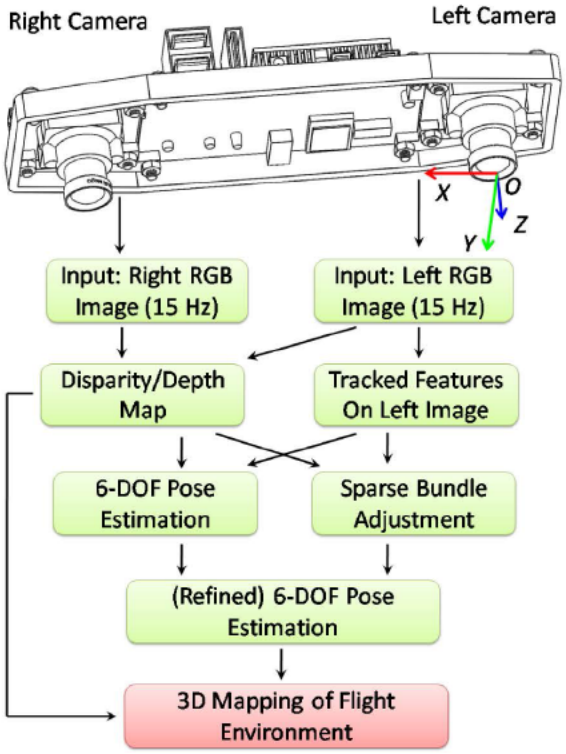


Fig. 4: Coordinate system $\{O\}$ of our stereo vision pre-processing system and whole software architecture.

We synchronized both cameras to publish the image pairs at 15 Hz with hardware trigger, and the image resolutions are 376×240. For the purpose of achieving real-time performance, we only estimate the disparity/depth map using *one* of two stereo image pairs instead of processing every consecutive image pairs, and the disparity/depth map is estimated based on an efficient stereo Semi-Global Matching (SGM) approach [9]. The disparity/depth map is generated in the optical frame of the left camera (i.e. reference camera, the coordinate system $\{O\}$ is located on its optical center). We assume that the disparity/depth map is estimated by the kth pair of stereo images, the coordinate of the ith feature with depth in $\{O^k\}$, i.e. $\mathbf{X}_i^k \in \mathbb{R}^3$, is

$$\begin{aligned}\mathbf{X}_i^k &= g(u_r^k, v_r^k, u_l^k, v_l^k) = [x_i^k, y_i^k, z_i^k]^T \\ &= \left[\frac{(u_l^k - c_x)b}{d^k}, \frac{(v_l^k - c_y)b}{d^k}, \frac{fb}{d^k}\right]^T \end{aligned} \quad (1)$$

i.e.

$$x_i^k = \frac{z_i^k(u_l^k - c_x)}{f}, y_i^k = \frac{z_i^k(v_l^k - c_y)}{f}$$

where, (u_l^k, v_l^k) and (u_r^k, v_r^k) are the pixels on the kth pair of left and right images, f represents the focal length, c_x and

[10]http://www.vision.caltech.edu/bouguetj/calib_doc/

c_y are the coordinates of optical center, $d^k = u_r^k - u_l^k$ is the disparity, and b is the baseline of the stereo vision system. Some estimated disparity/depth maps are shown in Fig. 5 and 9. Note: the (k+1)th pair of stereo images is *not* applied for estimating the disparity/depth map, i.e. $\mathbf{X}_i^{k+1}$=$z_i^{k+1}\hat{\mathbf{X}}_i^{k+1}$, where, $\hat{\mathbf{X}}_i^{k+1}$=$[\hat{x}_i^{k+1}, \hat{y}_i^{k+1}, 1]$.

B. Visual Feature Detection and Tracking

We extract and track the features on the consecutive reference (left) image parallelly using the *bucketing* [10] method, i.e. the input left image is divided into non-overlapped rectangle regions, and the maximum number of features to track in each bucket is set. Each bucket are smoothed with a Gaussian kernel to reduce noise firstly, then the FAST [11] detector is used to extract the keypoints, and a modified version of the BRIEF descriptor [12] is adopted to match the keypoints, the modified BRIEF descriptor of a keypoint $\mathbf{p}$, i.e. $D(\mathbf{p})$, is defined as:

$$D_j(\mathbf{p}) = \begin{cases} 1 & \text{if } \mathbf{p} + \mathbf{x}_j < \mathbf{p} + \mathbf{y}_j \\ 0 & \text{otherwise} \end{cases}, \forall j \in [1, ..., N_b] \quad (2)$$

where, $D_j(\mathbf{p})$ is the jth bit of the binary vector, $(\mathbf{x}_j, \mathbf{y}_j)$ is sampled in a $S_r \times S_r$ local neighbour region based on the locations of keypoints in the previous left image frame, however, $\mathbf{x}_j = \mathcal{N}(0, (\frac{1}{5}S_r)^2)$ and $\mathbf{y}_j = \mathcal{N}(\mathbf{x}_j, (\frac{2}{25}S_r)^2)$. The parameter N_b is the number of bits in the binary vector. The distance of two vectors is calculated by counting the number of different bits between them, i.e. Hamming distance, which is faster than computing the Euclidean distance.

C. 6-DOF Motion Estimation of Stereo Camera

In general, we defined $\mathbf{R} \in SO(3)$ as the rotation matrix, and $\mathbf{t} \in \mathbb{R}^3$ as the translation vector between two consecutive image frames in the left camera, then its motion can be defined as below:

$$\mathbf{X}_i^{k+1} = \mathbf{R} \cdot \mathbf{X}_i^k + \mathbf{t} \quad (3)$$

where, the $\mathbf{X}_i^k$ and $\mathbf{X}_i^{k+1}$ are the ith FAST feature coordinates in $\{O^k\}$ and $\{O^{k+1}\}$, respectively. However, a certain number of FAST features tracked by BRIEF descriptor only have 2D information because of the depth map estimation approach presented in Section III-A and the two bottlenecks of stereo camera introduced in Section I, i.e. $\mathbf{X}_i^{k+1} = z_i^{k+1}\hat{\mathbf{X}}_i^{k+1}$ and $\mathbf{X}_i^k = z_i^k\hat{\mathbf{X}}_i^k$.

Recently, J. Zhang et al [3] presented a promissing depth enhanced monocular odometry approach to effectively take advantage of 2D and 3D Harris corner [13] features tracked by the KLT tracker [14] in real-time motion estimation, which is demonstrated and utilized on the *RGB-D* sensor and *Monocular Camera-Lidar* device. In our work, we extended their motion estimation method for our newly designed stereo vision system. Then, a Bundle Adjustment (BA) [15] is adopted to refine the initially estimated camera motion parallelly, in which the keyframe is added when the tracked feature number is less than a threshold or camera motion is larger than a threshold.

D. 3D Mapping of UAV Flight Environment

Robust dynamic 3D mapping plays a key role for navigation/exploration and path planning during safely autonomous flights of UAV. In our work, the OctoMap[11] [16] is utilized as an efficient probabilistic 3D mapping framework for reconstructing arbitary UAV flight environments. This octree-based occupancy grid map models the occupied space (obstacles) and free areas clearly, and supports with coarse-to-fine resolutions. We applied a spherical coordinate system for representing map point, i.e. a map point is represented by its radial distance, polar angle and azimuthal angle. And the pre-processed infromation from our stereo vision preprocessing system are sent via a Gigabit Ethernet cable to the onboard host computer, i.e. Asctec Mastermind[12] with Intel i7-3612QE (4×2.1 GHz) and 4 GB DDR3 RAM, of UAV for real-time environment reconstruction and obstacle detection. Some 3D mapping results of UAV flight environments are shown in Fig. 7 and 10.

IV. SOFTWARE EVALUATION

In this section, the performance of the presented visual algorithm is evaluated using the well-known stereo video datasets from EuRoC[13] Challenge III project, which adopts the Asctec Firefly Hexcopter UAV to record stereo image pairs with Skybotix VI sensor, provides the Ground Truth (GT) of flying UAV pose and includes different dynamic environments with variant surrounding illuminations, blur motions and other challenging conditions.

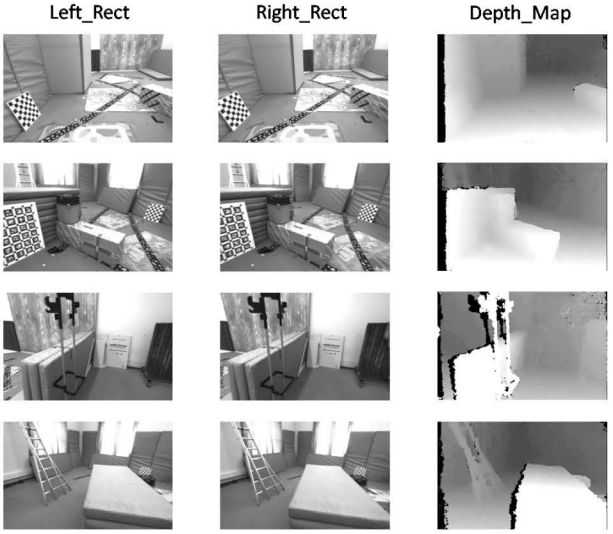


Fig. 5: Rectified stereo image pairs and estimated depth images captured during UAV flight. The localizations of images in row are shown in Fig. 7 with No.1, 2, 3 and 4.

For each input image frame, we divided it into 6×4 non-overlapped rectangle regions (i.e. buckets), and the maximum

[11] http://octomap.github.io/
[12] http://www.asctec.de
[13] www.euroc-project.eu/

number of FAST features in each bucket was set to 20, generating maximumly 480 FAST features to track in total. This method guarantees FAST features to evenly locate in each frame, enhancing the performance of pose estimation. Fig. 5, 6 and 7 show captured images, evaluation performance and 3D mapping result from the first challenging stereo video.

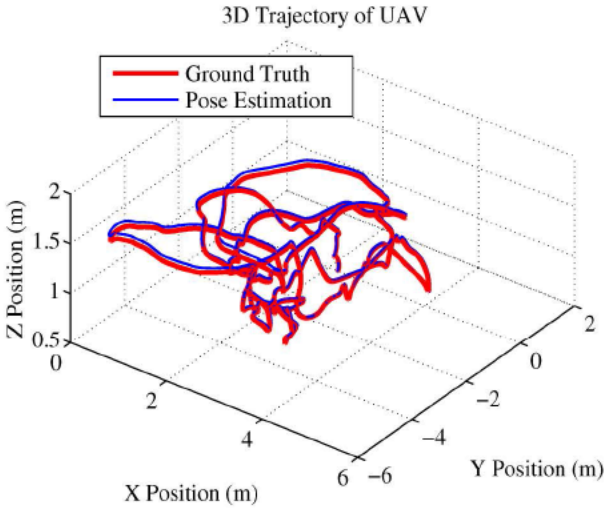


Fig. 6: Comparison of UAV 3D position estimation, the average 6D RMSE errors are shown in TABLE I.

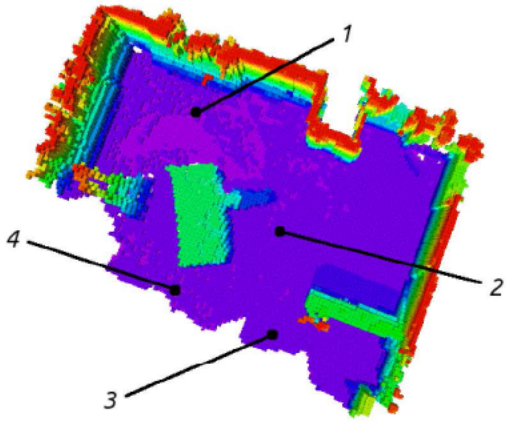


Fig. 7: 3D mapping result of UAV flight environment. Both voxel size and resolution are 0.1 meter. Red grids represent high altitudes, and purple grids indicate low altitudes.

TABLE I: The evaluation result. (Unit: Position error in mm, Orientation error in $degree$)

Parameter	X	Y	Z	Roll	Pitch	Yaw
RMSE	18.6	28.4	26.7	4.03	4.36	2.76

V. REAL FLIGHT TEST

After having evaluated our stereo visual algorithm, we mounted the newly designed light small-scale low-cost ARM-based stereo vision pre-processing system on the Asctec Pelican Quadrotor UAV, as shown in Fig. 8, to demonstrate the performance of visual algorithm with different types of dynamic GPS-denied environments. In this paper, one of the UAV flights in a corridor is presented, the approximate start and end locations in GPS coordinates are (40.439725, -3.689666) and (40.439861, -3.689026). The maximum flight speed of the UAV was ∼0.8m/s, with variant illuminations (e.g. floor/wall reflection) and blur motions as the main challenging problems during the UAV flight tests. Fig. 9 and 10 show captured images and 3D mapping result from the real UAV flight test.

Fig. 8: UAV flight test in a corridor environment.

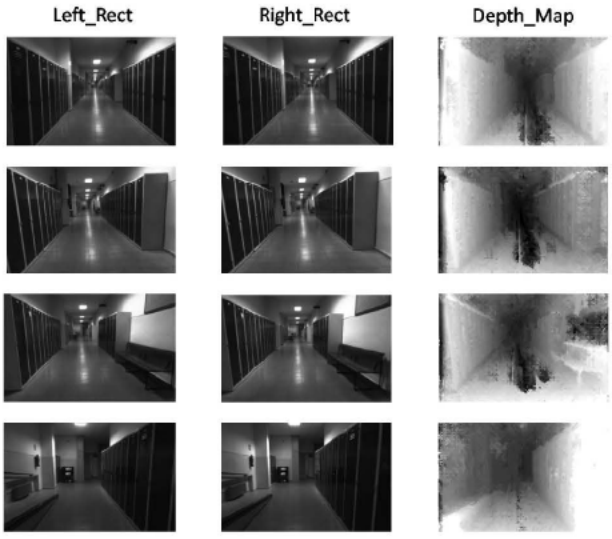


Fig. 9: Rectified stereo image pairs and estimated depth images captured in real UAV flight. The localizations of images in row are corresponding to the No.1, 2, 3 and 4 in Fig. 10.

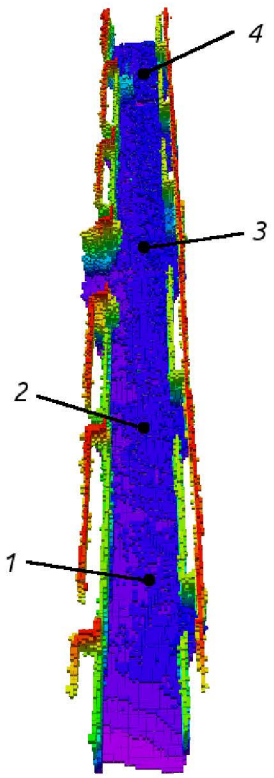


Fig. 10: 3D mapping result of real UAV flight test.

VI. CONCLUSIONS AND FUTURE WORKS

This paper proposed a newly designed light small-scale low-cost ARM-based stereo vision pre-processing system for typical small-scale/multi-rotor UAV. Compared to other existing state-of-art standalone stereo vision devices, our system has advantages in terms of size, weight, price, computation performance, mounting flexibility and the capability to process information from complex surrounding environments. Although this system only has a fixed small baseline, the presented software plugin was specifically developed for the system to effectively take advantage of those 2D and 3D features, thereby accurately estimating the 6D pose of UAV. Additionally, we have tested that the combination of FAST and BRIEF features guarantees the real-time performance of the stereo visual algorithm in the ARM architecture-based computer system. And OctoMap is applied as an efficient 3D occupancy grid mapping method to detect obstacles and reconstruct UAV flight environments. Finally, one of real UAV flight results is presented, which demonstrated the feasibility and robustness of our visual algorithm under the challenging situations due to illumination changes and blur motions.

For the future works, we will test our stereo vision system on different types of real small-scale/multi-rotor UAVs in large-scale outdoor GPS-denied cluttered dynamic environments, and finish keyframe-based global optimization algorithm in a loop-closure module to perform stereo SLAM. Moreover, the comparisons with other existing stereo VO/SLAM algorithms will be done. Finally, we also will add an IMU sensor to our stereo vision system to improve the pose estimation using a sensor fusion module.

ACKNOWLEDGMENT

The work reported in this paper is the consecution of several main research stages at the Universidad Politécnica de Madrid (UPM). The authors would like to thank the IRSES project within the Marie Curie Program FP7 - UECIMUAVS: USA and Europe Cooperation in Mini UAVs and the Spanish Ministry for Education, Culture and Sports for funding their international visitings. This work also has been sponsored by the Spanish Ministry of Science MICYT DPI 2010-20751-C02-01, MeSOANTEN project, TAISAP-UAV project and the China Scholarship Council (CSC).